# Jordan_PCA

*Jordan Chazin*

*March 8, 2016*

```
library(RNetCDF)
```

```
## Warning: package 'RNetCDF' was built under R version 3.2.3
```

```
# open and read file
fname<-"NOAA_Daily_phi_500mb.nc"
fid<-open.nc(fname)
print.nc(fid)
```

```
## dimensions:
##         T = 24873 ;
##         X = 144 ;
##         P = 1 ;
##         Y = 15 ;
## variables:
##         float T(T) ;
##                 T:standard_name = "time" ;
##                 T:pointwidth = 1 ;
##                 T:long_name = "Time" ;
##                 T:expires = 1454938800 ;
##                 T:calendar = "standard" ;
##                 T:gridtype = 0 ;
##                 T:units = "days since 1948-01-01 12:00:00" ;
##         float X(X) ;
##                 X:standard_name = "longitude" ;
##                 X:pointwidth = 2.5 ;
##                 X:long_name = "Longitude" ;
##                 X:gridtype = 1 ;
##                 X:units = "degree_east" ;
##         int P(P) ;
##                 P:long_name = "Pressure" ;
##                 P:gridtype = 0 ;
##                 P:units = "mb" ;
##         float Y(Y) ;
##                 Y:standard_name = "latitude" ;
##                 Y:pointwidth = 2.5 ;
##                 Y:long_name = "Latitude" ;
##                 Y:gridtype = 0 ;
##                 Y:units = "degree_north" ;
##         float phi(X, Y, P, T) ;
##                 phi:pointwidth = 0 ;
##                 phi:history = "T:  0000 1 Apr 2002 to 0000 5 Feb 2016 appended from datestring" ;
##                 phi:calendar = "standard" ;
##                 phi:center = "US Weather Service - National Met. Center" ;
##                 phi:gribparam = 7 ;
##                 phi:gribleveltype = 100 ;
##                  phi:gribvariable = 7 ;
##                  phi:PDS_TimeRange = 113 ;
##                  phi:process = "(180) 62 wave triangular, 28 layer Spectral model from "Medium Range
##                  phi:GRIBgridcode = 2 ;
##                  phi:gribNumBits = 9 ;
##                  phi:gribfield = 1 ;
##                  phi:subcenter = "NCEP Ensemble Products" ;
##                  phi:scale_min = -605 ;
##                  phi:grib_name = "HGT" ;
##                  phi:missing_value = 9.999e+20 ;
##                  phi:PTVersion = 2 ;
##                  phi:scale_max = 32480 ;
##                  phi:expires = 1454938800 ;
##                  phi:units = "gpm" ;
##                  phi:long_name = "Geopotential height" ;
##                  phi:standard_name = "geopotential_height" ;
```

```
data<-read.nc(fid)
close.nc(fid)

## verify dimensionality of data
## data$phi is matrix of pressure data organized
## on three axes: (lon, lat, days since 1/1/1948)
## e.g. last entry of matrix should be a singular
## pressure value:
head(data$phi[144,15,24873])
```

```
## [1] 5771
```

```
## for PCA, need to re-format into N x D matrix s.t.
## N = days
## D = lon x lat coördinate
## first create columns for D dimension:
ylat<-data$Y
xlon<-data$X

#reshape 144 x 15 x 24873 into NxD matrix of 24873x2160
phi.matrix <- t(matrix(data$phi,2160,24873))
dim(phi.matrix)
```

```
## [1] 24873  2160
```

```
# > dim(phi.matrix)
# [1] 24873  2160

# phi.matrix has 24,873 rows of daily pressure data
# the columns are ordered by lonXlat, and will be
# labeled as such in order to easily identify which
# columns "survive" the PCA.
# The labeling convention,
# will be as follows: the first columns
# will be labeled xlon[1]_x_ylat[1], xlon[2]_x_ylat[1],...
# and the last with xlon[143]_x_ylat[15], xlon[144]_x_ylat[15].
# Here I will label them to keep track.
xlon.factor <- as.factor(xlon)
ylat.factor <- as.factor(ylat)
a <- expand.grid(xlon.factor,ylat.factor)
a$coord <- paste(a$Var1,a$Var2,sep="_x_")
colnames(a)<-c("Lon","Lat","Lon_X_Lat")

colnames(phi.matrix) <- a$Lon_X_Lat

## Focusing on a Location and Time
# The only way to make PCA both intelligible
# and computationally feasible, is to focus on a
# window of time and a narrow geographic region.
# To be consistent with Phoebe and Hiroaki, I will
# do a PCA over the United States, for the months
# of June and July, 2015 (days 24624 to 24683)
# The longitude and latitude will be bounded by:

# Longitude: 230-300 Degrees East
# Latitude: 25-55 Degrees North

# these regular expressions help identify columsn which
# fit the Longitude and Latitude criteria
allCoords <- as.vector(a$Lon_X_Lat)
usa.regx.long <- grepl("^230_|^232.5_|^235_|^237.5_|^240_|^242.5_|^245_|^247.5_|^250_|^252.5_|^255_|^25
usa.regx.lat <- grepl("_25$|_27.5$|_30$|_32.5$|_35$|_37.5$|_40$|_42.5$|_45$|_47.5$|_50$|_52.5$|_55$",al
# Multiplying these vectors will yield a vector in which
# only "TRUE" fields are within USA's boundaries
usa.usa <- as.logical(usa.regx.lat*usa.regx.long)
# Now trim phi.matrix to June-July 2015, USA
phi.matrix <- phi.matrix[24624:24683,usa.usa]
dim(phi.matrix)
```

```
## [1]  60 261
```

```
# Now we need to center and scale each "grid", so that
# we can do more stable PCA:
## log transform
log.phi <- log(phi.matrix[,1:dim(phi.matrix)[2]])

## apply PCA with CENTERING and SCALING
phi.pca <- prcomp(log.phi, center = TRUE, scale. = TRUE, tol = .25)

# summary method
summary(phi.pca)
```

```
## Importance of components:
##                          PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     8.060 7.2453 5.8443 5.5795 4.2113 3.61319 3.21658
## Proportion of Variance 0.267 0.2158 0.1404 0.1280 0.0729 0.05366 0.04253
## Cumulative Proportion  0.267 0.4828 0.6232 0.7511 0.8240 0.87769 0.92022
##                             PC8     PC9    PC10
## Standard deviation     2.71364 2.59776 2.30177
## Proportion of Variance 0.03027 0.02774 0.02178
## Cumulative Proportion  0.95048 0.97822 1.00000
```

```r
# plot method
plot(phi.pca, type = "l")


# Try to produce a ggbiplot once PCA is complete
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 3.2.3
```

```r
install_github("vqv/ggbiplot")
```

```
## Skipping install for github remote, the SHA1 (7325e880) has not changed since last install.
##   Use `force = TRUE` to force installation
```

```r
library(ggbiplot)
```

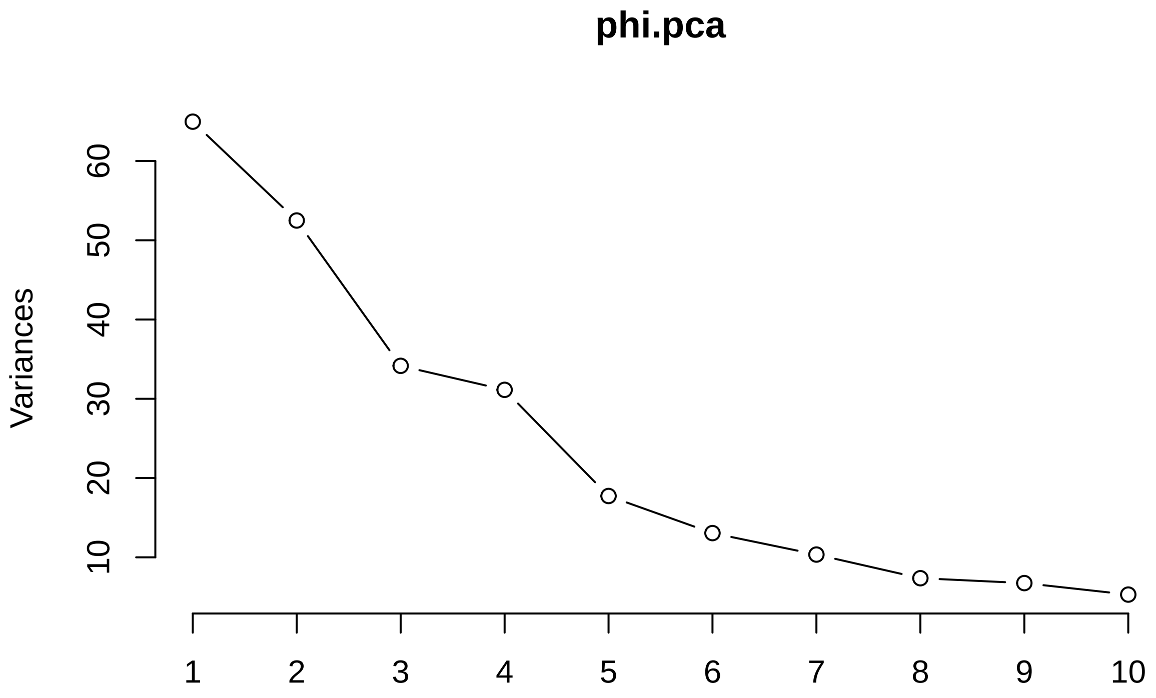
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

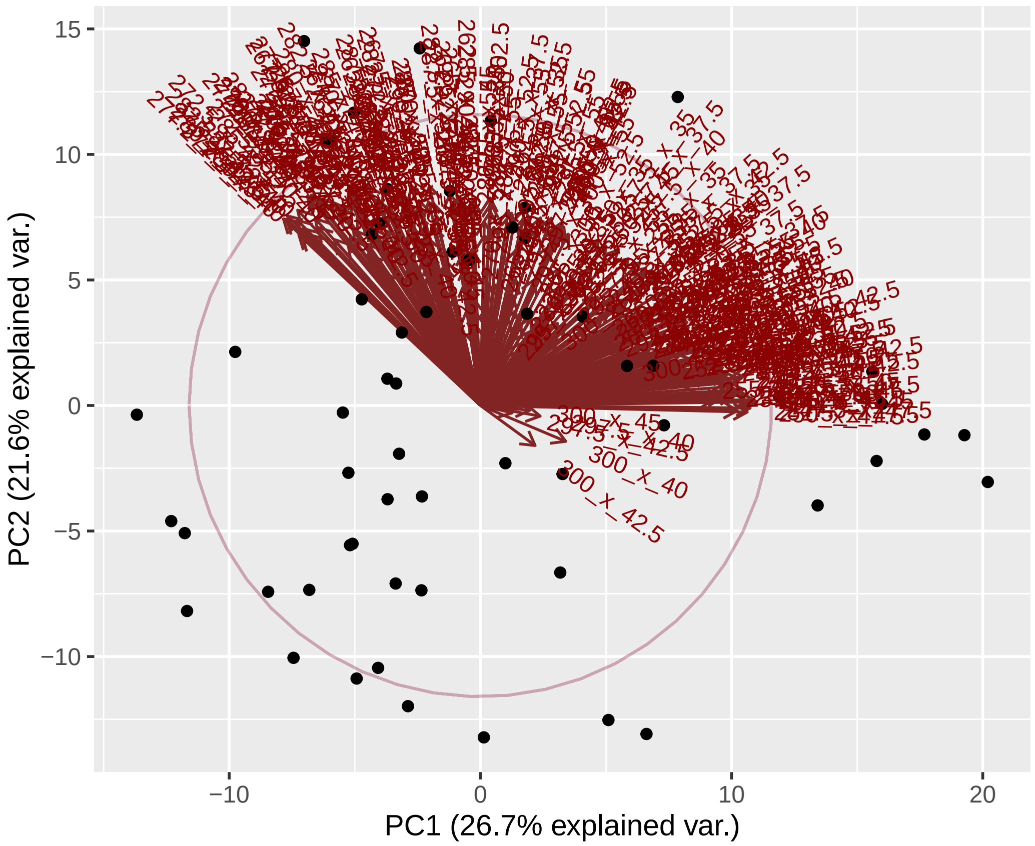
```
## Loading required package: plyr
## Loading required package: scales
```

```
## Warning: package 'scales' was built under R version 3.2.3
```

```
## Loading required package: grid
```



```r
g <- ggbiplot(phi.pca ,obs.scale = 1, var.scale = 1,
              ellipse = TRUE,
              circle = TRUE,alpha=1)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')
print(g)
```


PC2 (21.6% explained var.)
PC1 (26.7% explained var.)
300_x_45
300_x_40
300_x_42.5
300_x_40
300_x_42.5